PM & FM

Introduction

As we've discussed, amplitude modulation suffers from the problem that radio channels introduce their own amplitude modulation in the form of fading. You have seen in the labs that field amplitude can vary by 30 dB or more over a very short distance. So, amplitude is not a very good place to put information in a wireless channel, at least at higher frequencies. On the other hand, interference and fading does not affect radio frequency. Consequently putting information into the frequency, or phase, of an RF signal should be a more reliable way to communicate. Indeed, *frequency modulation* (FM), is the method of choice for quality analog radio communication. Because frequency is the derivative of phase, FM can be considered as a form of phase modulation (PM). PM is not of much use for analog communication but is important for digital radio systems. FM is useful for both analog and digital systems.

Phase modulation

As the name implies, in phase modulation we encode information by modulating the phase of the RF signal. The result is a signal of the form

$$s(t) = A_c \cos[\omega_c t + \phi(t)] \tag{20.1}$$

where the phase, relative to the RF carrier, is $\phi(t) = k_p m(t)$. Here $m(t)$ is the modulating signal and k_p is the *phase deviation constant.* If $m(t)$ is in volts, then k_p has units of radians per volt. Figure 20.1 shows an example of phase modulation.

A drawback to phase modulation is that phase is a relative concept, i.e., we need a reference to define phase relative to. It can be difficult to establish a common reference between a transmitter and receiver, especially for continuous (analog) modulation. Imagine that you are shown a sine wave on an oscilloscope and someone asks you "What is the phase of that waveform?" You can't answer unless you also have a reference sine wave and then determine the phase difference between the two. For example, in Fig. 20.1 the dotted waveform serves as a reference for the solid waveform. From the phase difference of these two you can determine the signal that produced the phase modulation. But you couldn't determine that solely from the solid waveform alone.

So, phase modulation is generally not very useful for analog modulation. Phase modulation is, however, very useful for digital modulation, and we will study that subject shortly. On the other hand, frequency, which is the derivative of phase, *is* an absolute quantity. You can look at a sinusoid on a scope and unambiguously determine its frequency without reference to any other signal.

$$f_a := 400 \quad f_r := 5000 \quad T := \frac{1}{f_a} \quad t := 0, \frac{T}{1000} .. T$$

$$\phi(t) := \pi \frac{\sin(2\pi f_a \cdot t) + \cos(4\pi f_a \cdot t)}{2} \qquad s(t) := \cos(2\pi f_r \cdot t + \phi(t))$$

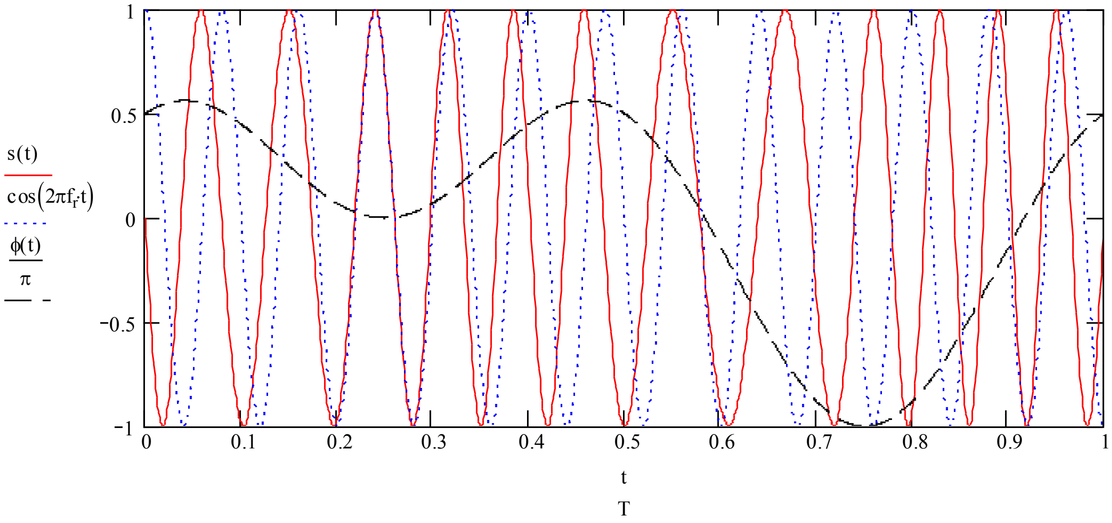

Figure 20.1:Mathcad phase modulation example. The “RF” carrier frequency is 5000. The modulating signal consists of sinusoids at frequencies of 400 and 800.

Frequency Modulation

Instantaneous frequency is proportional to the time derivative of phase. That is, given a signal $\cos\theta(t)$, the instantaneous frequency is

$$f(t) = \frac{1}{2\pi}\frac{d\theta}{dt} \tag{20.2}$$

The instantaneous frequency of a phase-modulated signal of the form (20.1) is

$$f(t) = f_c + \frac{1}{2\pi}\frac{d\phi}{dt} \tag{20.3}$$

that is, the carrier frequency plus a term proportional to the time derivative of the phase modulation. The idea of frequency modulation is to make $d\phi/dt$, instead of ϕ, proportional to the modulating signal. We set $d\phi/dt = 2\pi k_f m(t)$ so that

$$f(t) = f_c + k_f m(t) \tag{20.4}$$

where k_f is the *frequency deviation constant*. If $m(t)$ is in volts, then k_f has units of Hz per volt. For example, suppose k_f is 1000 Hz per volt and f_c is 100 MHz. Then to represent a signal of 2 volts we'd have our transmitter send a frequency of 100.002 MHz (100 MHz plus 2 kHz), to represent a signal of –3 volts we transmit 99.997 MHz and so on. Frequency is an absolute concept; if you see a sine wave on an oscilloscope you can determine its frequency by counting the number of peaks that occur in a given interval. Therefore no reference is required. FM can easily be generated with a voltage controlled oscillator (VCO), a device that generates an oscillation whose frequencies varies with an applied control voltage.

Since $d\phi/dt = 2\pi k_f m(t)$ we can write

$$\phi(t) = \phi(0) + 2\pi k_f \int_0^t m(x)dx \tag{20.5}$$

Taking $\phi(0) = 0$, (20.1) then gives us the RF signal

$$s(t) = A_c \cos\left(2\pi f_c t + 2\pi k_f \int_0^t m(x)dx\right) \tag{20.6}$$

This looks pretty ugly due to the integral/derivative relationship between phase and frequency, but keep in mind that what's really going on is the simple frequency shifting described in (20.4). An example of frequency modulation is shown in Fig. 20.2.

Demodulating an received FM signal is, in principle, easy. A traditional approach is to use a circuit with a response that varies more-or-less linearly with frequency. For example, an ideal differentiator has a response $H(\omega) = j\omega$, so the amplitude of the response is proportional to frequency. This converts frequency into amplitude, i.e., FM into AM. Then an envelope detector can recover the original modulation. A more digital approach to FM reception is to use a counter that acts like a frequency meter and determines how many oscillations occur during a short time interval.

$$f_a := 400 \qquad f_r := 5000 \quad T := \frac{1}{f_a} \qquad t := 0, \frac{T}{1000} .. T \qquad\qquad k_f := 1000$$

$$m(t) := \frac{\sin(2\pi\, f_a \cdot t) + \cos(4\pi\, f_a \cdot t)}{2} \qquad\qquad s(t) := \cos\left(2\pi\, f_r \cdot t + 2\pi\, k_f \cdot \int_0^t m(x)\, dx\right)$$

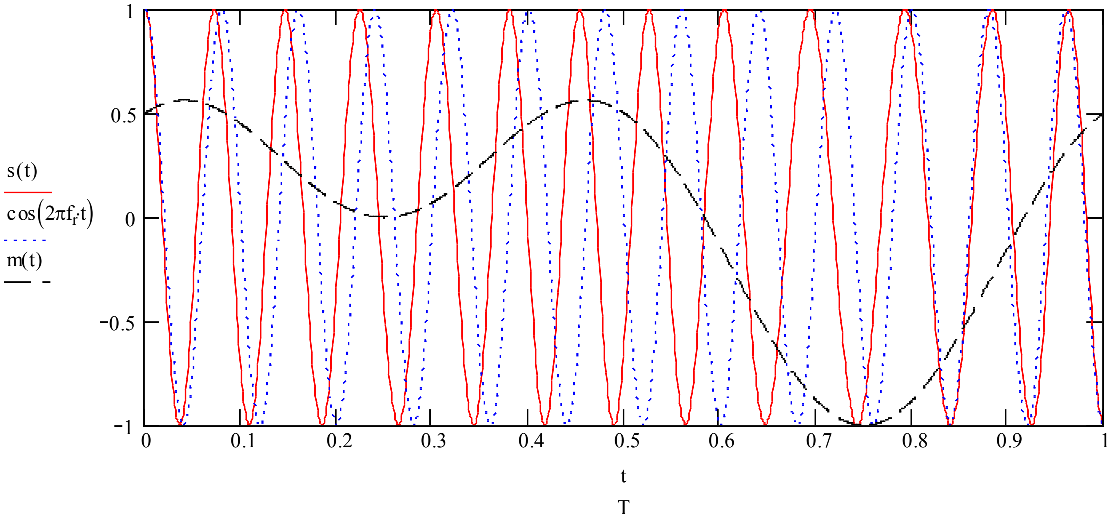

Figure 20.2: Mathcad frequency modulation example. The "RF" carrier and modulating signal are the same as in Fig. 20.1.

Spectrum of an FM Signal

For an arbitrary $m(t)$, expression (20.6) can get pretty messy. Let's limit ourselves to a sinusoidal test tone $m(t) = a_m \cos 2\pi f_m t$. The phase is

$$\begin{aligned} \phi(t) &= 2\pi k_f a_m \int_0^t \cos(2\pi f_m x)dx \\ &= \frac{a_m k_f}{f_m} \sin(2\pi f_m t) \\ &= \frac{\Delta f}{f_m} \sin(2\pi f_m t) \end{aligned} \tag{20.7}$$

So the RF signal is

$$s(t) = A_c \cos\left(2\pi f_c t + \beta \sin(2\pi f_m t)\right) \tag{20.8}$$

where $\Delta f = a_m k_f$ is the *frequency deviation*, and $\beta = \Delta f / f_m$ is the *frequency modulation index.* For example, if a_m is 0.1 volts, and k_f is 10 kHz per volt, then the frequency deviation is 1 kHz.

This means that the instantaneous frequency varies $\pm 1\text{kHz}$ about the carrier frequency. You get the same signal if you have an a_m of 1 volt and a k_f of 1 kHz per volt; it's the product that is important. Physically this would mean you could drive a high-sensitivity (large k_f) VCO with a small signal, or drive a low-sensitivity VCO with a large signal and you would get the same result. In practice, therefore, it is Δf that is specified. For example, broadcast FM radio has a specified frequency deviation of 75 kHz.

The frequency modulation index is dimensionless and is the ratio of the frequency deviation to the modulation frequency. As the sine factor in (20.8) varies between ± 1, the phase varies over $\pm\beta$. As we'll see below β is important in determining the bandwidth of the RF signal.

Even for the simple case of sinusoidal modulation, the FM signal (20.8) is not easy to analyze. However, it is a periodic signal in the sense that for $f_m t = 0,1,2,3,\dots$ the phase term repeats. So, we should be able to represent it as a Fourier series. You can do this but the math is messy and requires some theory of Bessel functions. You find

$$\begin{aligned}\cos\left(2\pi f_c t + \beta\sin(2\pi f_m t)\right) = \; & J_0(\beta)\cos 2\pi f_c t \\ & + J_1(\beta)\left[\cos 2\pi(f_c + f_m)t - \cos 2\pi(f_c - f_m)t\right] \\ & + J_2(\beta)\left[\cos 2\pi(f_c + 2f_m)t + \cos 2\pi(f_c - 2f_m)t\right] \\ & + J_3(\beta)\left[\cos 2\pi(f_c + 3f_m)t - \cos 2\pi(f_c - 3f_m)t\right] \\ & + J_4(\beta)\left[\cos 2\pi(f_c + 4f_m)t + \cos 2\pi(f_c - 4f_m)t\right] \\ & + \cdots \end{aligned} \tag{20.9}$$

where $J_n(\beta)$ is the Bessel function of order n. The Bessel functions have the following power-series representations

$$J_n(\beta) = \left(\frac{\beta}{2}\right)^n \left[\frac{1}{0!n!} - \frac{1}{1!(n+1)!}\left(\frac{\beta}{2}\right)^2 + \frac{1}{2!(n+2)!}\left(\frac{\beta}{2}\right)^4 - \cdots\right] \tag{20.10}$$

The first four of these is plotted vs. β in Fig. 20.3.

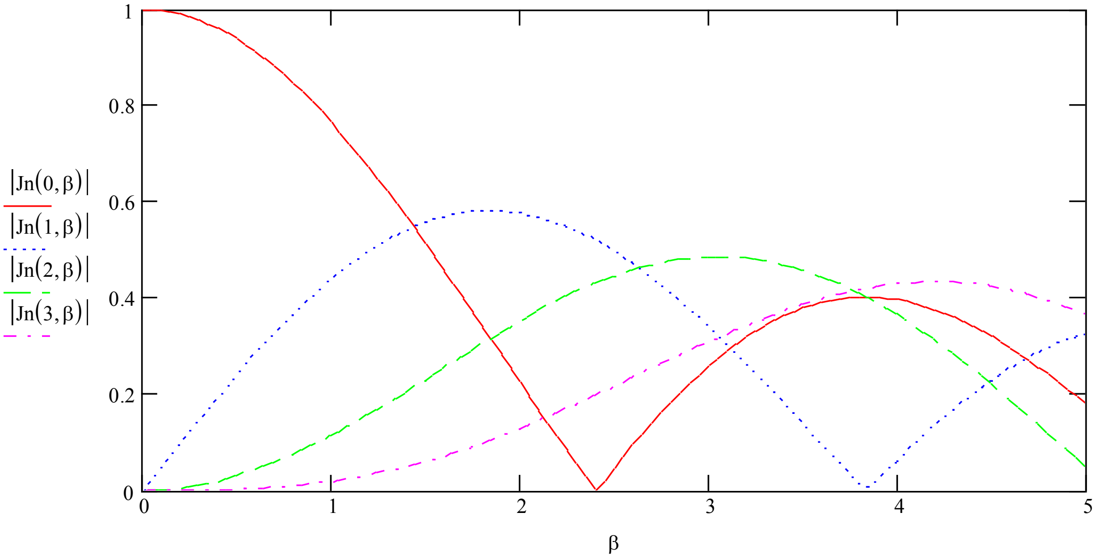

Figure 20.3: The first four Bessel functions ($n = 0,1,2,3$) vs. the frequency modulation index.

You can see from Fig. 20.3 that the n^{th} Bessel function more-or-less "turns on" at about $\beta = n$. So, roughly speaking, we will get a significant contribution from the terms in (20.10) up to $n \approx \beta$ for which the Fourier series terms have frequencies of $f_c \pm nf_m$.

Unlike the case of an AM signal, there is no finite spectral width into which all the signal power falls because the series (20.9) extends out to infinite frequency. However, due to the "turning on" property of the Bessel functions, nearly all the signal power will fall into a finite bandwidth. *Carson's rule* approximates the bandwidth of an FM signal as

$$B = 2(\beta + 1)f_m \tag{20.11}$$

If the modulating signal is more complex than a single sinusoid, f_m is taken as the maximum frequency. For example, phone quality audio typically includes frequencies up to about 4 kHz. So for this type of audio modulation we'd take $B = 2(\beta + 1)(4\,\text{kHz})$.

Note that $\beta f_m = \Delta f$, so we can write

$$B = 2\Delta f + 2f_m \tag{20.12}$$

The bandwidth is about twice the frequency deviation plus twice the maximum frequency in the modulating signal. An example of the spectrum of an FM radio signal is shown in Fig. 20.4

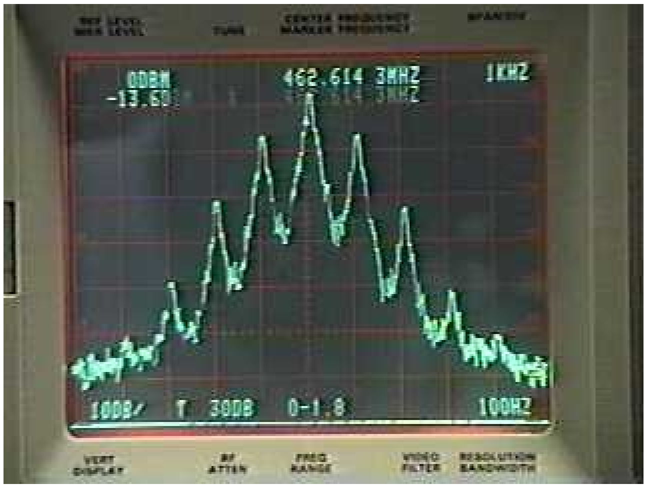

Figure 20.4: Spectrum of a 462-MHz FM radio transmission with 1kHz modulation.

Note that unlike the case for AM where the modulation index α could not exceed 1, the frequency modulation index β can, in theory, be arbitrarily large. For AM, varying the modulation index did not change the RF bandwidth, but for FM it does. The larger β is the larger our RF bandwidth is. So why not make β very small? You might guess that it has to do with *S/N*. Recall that for AM, *S/N* is proportional to α^2. We'll see that for FM, *S/N* is proportional to β^2. However, unlike the AM case, with FM we can have an arbitrarily large modulation index. Therefore, FM will allows us to trade off an increases bandwidth for an increasing *S/N*.

Before we discuss *S/N*, however, let's think about the effects of small-scale fading that we noted make AM impractical at higher frequencies. For FM we don't care about the amplitude. The carrier amplitude A_c can fluctuate by orders of magnitude and it won't change the phase/frequency of the signal. In fact, FM receivers typically employ a *limiter* that amplifies/compresses the signal so it has constant amplitude before performing demodulation. Fading can have an effect, however, on the *S/N* ratio and hence on the quality of the demodulated signal.

Effect of Noise in PM and FM Systems

In phase or frequency modulation, the signal is given by

$$s(t) = A_c \cos[\omega_c t + \phi(t)] \tag{12.13}$$

where either ϕ or its time derivative is proportional to the modulating signal $m(t)$. As we did for AM, we'll represent the noise has having two components, one in phase with the carrier and one out of phase:

$$\begin{aligned} n(t) &= [n_c(t)\cos\omega_c t - n_s(t)\sin\omega_c t] \\ &= r_n(t)\cos[\omega_c t + \phi_n(t)] \end{aligned} \tag{12.14}$$

The two functions $n_c(t), n_s(t)$ are assumed to be zero-mean, Gaussian RVs with variance $\sigma_n{}^2$. The second form, which represents the noise in terms of random amplitude and phase modulation, is more convenient for us in the present case. You can show that $r_n(t)$ is Rayleigh distributed with $\left\langle r_n(t)^2 \right\rangle = 2\sigma_n{}^2$ while $\phi_n(t)$ is uniformly distributed between 0 and 2π. If we draw the signal and noise on a phasor diagram, we get the following picture.

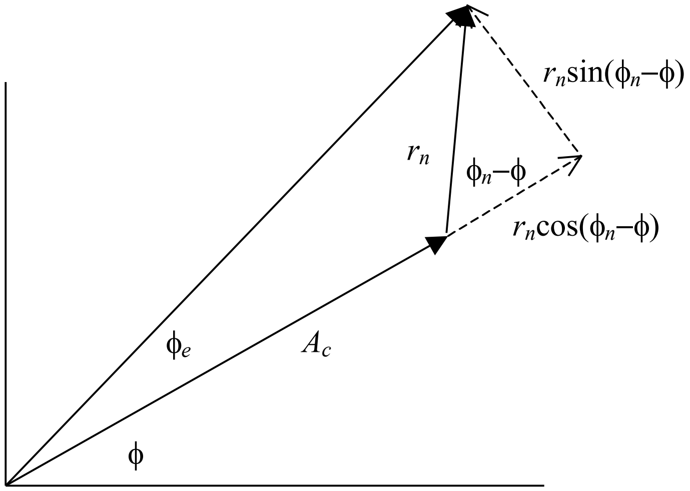

Figure 20.5: Signal and noise phasor diagram for PM/FM.

The signal has amplitude of A_c and a phase of $\phi(t)$. The noise phasor adds to that with amplitude of $r_n(t)$ and a phase relative to the signal of $\phi_n(t) - \phi(t)$. The phase of the resulting sum will differ from the signal phase by an amount $\phi_e(t)$. Trigonometry gives us

$$\phi_e(t) = \tan^{-1} \frac{r_n(t)\sin[\phi_n(t) - \phi(t)]}{A_c + r_n(t)\cos[\phi_n(t) - \phi(t)]} \tag{12.15}$$

if $A_c >> \sigma_n$ then almost always $A_c >> r_n(t)$ and the second term in the denominator can be neglected. Also, the angle $\phi_e(t)$ will be small so that $\tan^{-1}\phi \approx \phi$ and our recovered phase signal is

$$\phi_r(t) = \phi(t) + \frac{1}{A_c} r_n(t)\sin[\phi_n(t) - \phi(t)] \tag{12.16}$$

Recall that for FM $m(t) = \frac{1}{2\pi k_f}\frac{d}{dt}\phi(t)$. Therefore

$$m_r(t) = m(t) + \frac{1}{2\pi k_f A_c}\frac{d}{dt}\{r_n(t)\sin[\phi_n(t) - \phi(t)]\} \tag{12.17}$$

This is similar to (19.15) for the AM case, however, there is a *derivative* of the noise term here. Since derivation in the time domain corresponds to multiplication by $j\omega$ in the frequency domain, we can see that the derivative will have an amplitude that increases with f_m. We'll end up with a factor of $f_m / k_f \propto 1/\beta$. With the audio noise amplitude decreased by this factor, the noise power will be decreased by the square of it, and as a result the *S/N* of the recovered modulation should be proportional to β^2. More detailed analysis shows that

$$S / N_{audio,FM} = 3\beta^2 S / N_{audio,AM} \tag{12.18}$$

where $S / N_{audio,AM}$ is the audio *S/N* for an AM receiver with the same signal power and noise spectral density, and using 100% amplitude modulation. Since there is no theoretical limit on how large β can be, we can, in theory, increase *S/N* arbitrarily by increasing the bandwidth of the RF signal, $2\Delta f + 2f_m$. For example, broadcast FM specifies $f_m = 15\,\text{kHz}$ and $\Delta f = 75\,\text{kHz}$. So, $\beta = 5$ and the audio *S/N* is 75 times (about 19 dB) what you'd get from an AM station at the same received power level.

If the *S/N* is very low so that $A_c << \sigma_n$, then we could redraw Fig. 20.5, exchanging $\phi(t)$ and $\phi_n(t)$ and exchanging A_c and $r_n(t)$, to get

$$\phi_r(t) = \phi_n(t) + \frac{A_c}{r_n(t)} \sin\left[\phi(t) - \phi_n(t)\right] \tag{12.19}$$

The only place the signal $\phi(t)$ appears is inside the sine function where it is added to the uniformly distributed random phase $-\phi_n(t)$. The result is that the signal as effectively disappeared. As in the AM case, we see that FM exhibits a *threshold effect.* For this reason the *S/N* gains implied by (12.18) have a limit. If the RF *S/N* is too low to begin with (much below about 10 dB, say) then you end up in the "below threshold" situation and you have no signal no matter what the value of β.

References

1. Ziemer, R. E. and W. H. Tranter, *Principles of Communcations*, Houghton Mifflin, 1985, ISBN 0-395-35724-1.

2. Proakis, J. G. and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice Hall, 2002, ISBN 0-13-061793-8.

3. Couch, L. W., *Digital and Analog Communication Systems*, 6th ed., Prentice Hall, 2001, ISBN 0-13-081223-4.